What is claimed is:

1. A method for controlling a web farm having a plurality of websites and servers, the method comprising:

categorizing customer requests received from said websites into a plurality of categories, said categories comprising a shareable customer requests and unshareable customer requests;

routing said shareable customer requests such that any of said servers may process shareable customer requests received from different said websites; and

routing said unshareable customer requests from specific said websites only to specific servers to which said specific websites have been assigned.

2. The method of claim 1 further comprising a Goal procedure, said Goal procedure comprising determining, for each said customer request, an optimal server from among said servers to which each said customer request is to be assigned so as to minimize an average customer response time at any given moment, given said assignment of said websites to said servers and a current customer request load.

3. The method of claim 2 wherein said Goal procedure is effected by minimizing the function

$$\sum_{j=1}^{N} R_j\left(\sum_{i=1}^{M}\left(x_{i,j}+y_{i,j}\right)\right)$$

is minimized subject to the constraints

$$\sum\left(x_{i,j}+y_{i,j}\right)\in\{0,\ldots,L_j\},$$

$$\sum_{j=1}^{N} x_{i,j}=c_i,$$

$$x_{ij}=0 \text{ if } a_{ij}=0,$$

$$\sum_{j=1}^{N} y_{i,j}=d_i, \text{ and}$$

$$y_{ij}=0 \text{ if } b_{ij}=0,$$

where M is the number of websites, N is the number of servers, R_j is the expected response time as a function of customer arrival rate at server j, x_{ij} is a decision variable representing the hypothetical

number of shareable requests for website i that might be handled by server j, y_{ij} is a decision variable representing the hypothetical number of unshareable requests for website i that might be handled by server j, L_j is the maximum acceptable load for server j, c_i is the current number of shareable customer requests from website i, d_i is the current number of unshareable requests from website i, $a_{i,j}$ is an index indicating if shareable requests from website i may be routed to server j, and $b_{i,j}$ is an index indicating if unshareable requests from website i may be routed to server j.

4. The method of claim 3 further comprising

creating and maintaining a directed graph, said directed graph comprising a dummy node and a plurality of server nodes, each said server node corresponding to one of said servers;

designating one of said sever nodes a winning node for which the expression

$$R_j\left(\sum_{i=1}^{M}\left(x_{i,j}+y_{i,j}+1\right)\right)-R_j\left(\sum_{i=1}^{M}\left(x_{i,J}+y_{i,J}\right)\right)$$

is minimal; and

choosing a shortest directed path from said dummy node to said winning node.

5. The method of claim 1 further comprising a Static procedure, said Static procedure comprising assigning specific said websites to specific said servers for the purposes of processing unsharable customer requests.

6. The method of claim 5 wherein said Static procedure assigns said websites to specific servers based upon forecasted demand for shareable and unsharable customer requests from each said website.

7. The method of claim 2 further comprising a Dynamic procedure, said Dynamic procedure comprising:

examining the next customer request;

invoking said Goal procedure in order to determine which server is the optimal server to currently process said next customer request; and

dispatching said next customer request to said optimal server.

8. The method of claim 7 further comprising:

receiving said customer requests into a queue; and

wherein said Dynamic procedure further comprises:

monitoring said customer requests in said queue;

monitoring customer requests currently being processed by said servers;

defining, for each j^{th} server, a function $\dot{R}_j(z)$ by setting

$$\dot{R}_j(z) = R_j\left(z + \sum\left(\ddot{c}_{i,j} + \ddot{d}_{i,j}\right)\right);$$

defining, for each j^{th} server, a revised acceptable load limit $\dot{L}_j$ by setting

$$\dot{L}_j = L_j - \sum_{i-1}^{M}\left(\ddot{c}_{i,j} + \ddot{d}_{i,j}\right); \text{ and}$$

invoking said Goal procedure to utilize said $\dot{R}_j(z)$ function and revised acceptable load limit $\dot{L}_j$ to minimize the function

$$\sum_{j=1}^{N}\dot{R}_j\left(\sum_{i=1}^{M}\left(\dot{x}_{i,j} + \dot{y}_{i,j}\right)\right)$$

subject to the constraints:

$$\sum_{i=1}^{M}\left(\dot{x}_{i,j} + \dot{y}_{i,j}\right) \in \{0, \ldots, \dot{L}_j\},$$

$$\sum_{j=1}^{N}\dot{x}_{i,j} = \dot{c}_i\,,$$

$$\dot{x}_{i,j} = 0 \text{ if } a_{i,j} = 0\,,$$

$$\sum_{j=1}^{N}\dot{y}_{i,J} = \dot{d}_i\,, \text{ and}$$

$$\dot{y}_{i,J} = 0 \text{ if } b_{i,j} = 0\,,$$

where $\dot{x}_{i,j}$ is a decision variable representing the hypothetical number of shareable requests in the queue for website i that might be handled by server j, $\dot{y}_{i,j}$ is a decision variable representing the hypothetical number of unshareable requests for website i that might be handled by server j, $\dot{c}_i$ is the current number of shareable customer requests in the queue from website i, $\dot{d}_i$ is the current

number of unshareable requests in the queue from website i, $\ddot{c}_i$ is the current number of shareable customer requests from website i currently being processed in one of the servers, and $\ddot{d}_i$ is the current number of unshareable requests from website i currently being processed in one of the servers.

9. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for controlling a web farm having a plurality of websites and servers, said method steps comprising:

categorizing customer requests received from said websites into a plurality of categories, said categories comprising a shareable customer requests and unshareable customer requests;

routing said shareable customer requests such that any of said servers may process shareable customer requests received from different said websites; and

routing said unshareable customer requests from specific said websites only to specific servers to which said specific websites have been assigned.

10. The apparatus of claim 9 further comprising a Goal procedure, said Goal procedure comprising determining, for each said customer request, an optimal server from among said servers to which each said customer request is to be assigned so as to minimize an average customer response time at any given moment, given said assignment of said websites to said servers and a current customer request load.

11. The apparatus of claim 10 wherein said Goal procedure is effected by minimizing the function

$$\sum_{j=1}^{N} R_j \left(\sum_{i=1}^{M} \left(x_{i,j} + y_{i,j} \right) \right)$$

is minimized subject to the constraints

$$\sum \left(x_{i,j} + y_{i,j} \right) \in \{0, \ldots, L_j\}$$

,

$$\sum_{j=1}^{N} x_{i,j} = c_i$$

,

$$x_{i,j} = 0 \text{ if } a_{i,j} = 0,$$

$$\sum_{j=1}^{N} y_{i,j} = d_i \quad \text{, and}$$

$$y_{i,j} = 0 \text{ if } b_{i,j} = 0,$$

where M is the number of websites, N is the number of servers, R_j is the expected response time as a function of customer arrival rate at server j, $x_{i,j}$ is a decision variable representing the hypothetical number of shareable requests for website i that might be handled by server j, $y_{i,j}$ is a decision variable representing the hypothetical number of unshareable requests for website i that might be handled by server j, L_j is the maximum acceptable load for server j, c_i is the current number of shareable customer requests from website i, d_i is the current number of unshareable requests from website i, $a_{i,j}$ is an index indicating if shareable requests from website i may be routed to server j, and $b_{i,j}$ is an index indicating if unshareable requests from website i may be routed to server j.

12. The apparatus of claim 11 further comprising

creating and maintaining a directed graph, said directed graph comprising a dummy node and a plurality of server nodes, each said server node corresponding to one of said servers;

designating one of said sever nodes a winning node for which the expression

$$R_j\left(\sum_{i=1}^{M}\left(x_{i,j} + y_{i,j} + 1\right)\right) - R_j\left(\sum_{i=1}^{M}\left(x_{i,j} + y_{i,j}\right)\right)$$

is minimal; and

choosing a shortest directed path from said dummy node to said winning node.

13. The apparatus of claim 9 further comprising a Static procedure, said Static procedure comprising assigning specific said websites to specific said servers.

14. The apparatus of claim 13 wherein said Static procedure assigns said websites to specific servers based upon forecasted demand for shareable and unsharable customer requests from each said website.

15. The apparatus of claim 10 further comprising a Dynamic procedure, said Dynamic procedure comprising:

examining the next customer request;

invoking said Goal procedure in order to determine which server is the optimal server to currently process said next customer request; and

dispatching said next customer request to said optimal server.

16. The apparatus of claim 15 further comprising:

receiving said customer requests into a queue; and

wherein said Dynamic procedure further comprises:

monitoring said customer requests in said queue;

monitoring customer requests currently being processed by said servers;

defining, for each j^{th} server, a function $\dot{R}_j(z)$ by setting

$$\dot{R}_j(z) = R_j\left(z + \sum\left(\ddot{c}_{i,j} + \ddot{d}_{i,j}\right)\right);$$

defining, for each j^{th} server, a revised acceptable load limit $\dot{L}_j$ by setting

$$\dot{L}_j = L_j - \sum_{i-1}^{M}\left(\ddot{c}_{i,j} + \ddot{d}_{i,j}\right); \text{ and}$$

invoking said Goal procedure to utilize said $\dot{R}_j(z)$ function and revised acceptable load limit $\dot{L}_j$ to minimize the function

$$\sum_{j=1}^{N}\dot{R}_j\left(\sum_{i=1}^{M}\left(\dot{x}_{i,j} + \dot{y}_{i,j}\right)\right)$$

subject to the constraints:

$$\sum_{i=1}^{M}\left(\dot{x}_{i,j} + \dot{y}_{i,j}\right) \in \{0, \ldots, \dot{L}_j\},$$

$$\sum_{j=1}^{N}\dot{x}_{i,j} = \dot{c}_i\,,$$

$$\dot{x}_{i,j} = 0 \text{ if } a_{i,j} = 0,$$

$$\sum_{J=1}^{N} \dot{y}_{i,J} = \dot{d}_i \text{, and}$$

$$\dot{y}_{i,j} = 0 \text{ if } b_{i,j} = 0,$$

where $\dot{x}_{i,j}$ is a decision variable representing the hypothetical number of shareable requests in the queue for website i that might be handled by server j, $\dot{y}_{i,j}$ is a decision variable representing the hypothetical number of unshareable requests for website i that might be handled by server j, $\dot{c}_i$ is the current number of shareable customer requests in the queue from website i, $\dot{d}_i$ is the current number of unshareable requests in the queue from website i, $\ddot{c}_i$ is the current number of shareable customer requests from website i currently being processed in one of the servers, and $\ddot{d}_i$ is the current number of unshareable requests from website i currently being processed in one of the servers.

17. A web farm, comprising:

means for receiving customer requests from customer;

means for processing said customer requests to produce responses;

means for transmitting said responses to said customers;

means for categorizing said customer requests into shareable customer requests and unshareable customer requests;

a network dispatcher comprising means for executing a Goal procedure, a Static procedure, and a Dynamic procedure;

said Goal procedure comprising determining, for each said customer request, an optimal server from among said servers to which each said customer request is to be assigned so as to minimize an average customer response time at any given moment, given said assignment of said websites to said servers and a current customer request load, wherein said shareable customer requests may be assigned to any said server and wherein said unshareable customer requests may only be assigned to specific servers depending on which said website said unshareable customer request originated;

said Static procedure comprising assigning specific said websites to specific said servers; and

said Dynamic procedure comprising:

examining the next customer request;

invoking said Goal procedure in order to determine which server is the optimal server to currently process said next customer request; and

dispatching said next customer request to said optimal server.